ED 137 365 95 TM 006 166

AUTHOR St. Pierre, Robert G.; Ladner, Rosamund
TITLE Correcting Covariates for Unreliability: Does It Lead to Differences in an Evaluator's Conclusions?
INSTITUTION Abt Associates, Inc. Cambridge, Mass.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE [Apr 77]
CONTRACT 300-75-0134
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Achievement Tests; *Analysis of Covariance; Compensatory Education Programs; Early Childhood Education; *Program Evaluation; *Test Reliability; *True Scores
IDENTIFIERS Project Follow Through

ABSTRACT

One specific correction model suggested by Cohen and Cohen (1975) is applied to data collected in the evaluation of a large-scale quasi-experimental program (Project Follow Through), and the effects of different assumptions about test reliability on the analysis results and on the conclusions of the evaluators are examined. The study determines whether the application of reliability or "true score" corrections alters the results obtained via an analysis employing uncorrected covariates in such a fashion as to appreciably change the policy-oriented conclusions of an evaluator. The data on which this paper is based were collected for the 1976 Follow Through evaluation and include measures on a total of over 5,000 children who began the program in kindergarten (Fall 1971) and completed it in third-grade (Spring 1975). Results indicate that application of true-score corrections using three separate reliability estimates to covariates employed on analysis of covariance did not change the conclusions of the Follow Through evaluators. (RC)

CORRECTING COVARIATES FOR UNRELIABILITY: DOES IT LEAD TO DIFFERENCES IN AN EVALUATOR'S CONCLUSIONS?

BY

ROBERT G. ST.PIERRE
AND
ROSAMUND LADNER

ABT ASSOCIATES INC.
55 Wheeler Street
Cambridge, MA 02138

Presented at the Annual Meetings of the American Educational Research Association, New York, April 4-8, 1977.

This paper is based on research performed by Abt Associates Inc. under Contract No. 300-75-0134 to the United States Office of Education. Neither Abt Associates staff nor the Office of Education is responsible for any errors or omissions: the authors accept full responsibility for those. The statements in this paper do not necessarily represent any official position of the Office of Education.

Among the many problems prevalent in the evaluation of educational programs are those concerned with the adjustment of outcome scores based on one or more covariates. Typically, evaluations of these programs are implemented in a quasi-experimental fashion, and some version of the analysis of covariance (ANCOVA) is employed in an attempt to statistically equate treatment and comparison groups on one or more pretreatment conditions. However, the application of ANCOVA to quasi-experimental data has been widely criticized because violation of the assumption that subjects should be randomly assigned to treatment and comparison groups leads to systematic bias (usually underadjustment when the treatment group is initially disadvantaged with respect to the control group) of outcome scores (Campbell and Boruch, 1975). Achievement tests are commonly used as outcome measures for educational programs. Also, they are often employed as premeasures and serve as covariates in subsequent analyses. Since such tests are known to contain error, it has been argued that they should be corrected for unreliability prior to entry into a covariance analysis (Lord, 1960).

The current debate about the merits of correction for unreliability has raised many methodological questions. For example, which of a wide variety of correction formulas should be used, and which of many available estimates of test reliability is appropriate? This paper does not add to or review the methodological literature, but instead applies one specific correction model suggested by Cohen and Cohen (1975) to data collected in the evaluation of a large-scale quasi-experimental program (Project Follow Through), and examines the effects of different assumptions about test reliability on the analysis results and on the conclusions of the evaluators. The purpose of the study is to determine whether the application of reliability or "true-score" corrections alters the results obtained via an analysis employing uncorrected covariates in such a fashion as to appreciably change the policy-oriented conclusions of an evaluator.

Background

The origins of Follow Through can be traced to an early evaluation of Project Head Start (Wolff and Stein, 1966) which asserted that the 1965 Head Start experiences had increased the Head Start children's school readiness. The fact that these presumed increases were not reflected in the achievement test performance of the children at the end of their kindergarten experience in 1966 was attributed to the inappropriateness of traditional elementary education. Although some critics viewed this study as raising questions about the value of Head Start, the Johnson administration proposed a Follow Through program which would continue service to disadvantaged children through third grade. Funding problems forced a change in the emphasis of Follow Through from a full-scale service program to an experimental program in education in which educational specialists (sponsors) sponsored a variety of educational models in groups of school districts (sites). The educational strategies included: highly structured projects emphasizing academic skills in reading and arithmetic; projects stressing cognitive thinking through asking and answering questions, problem solving, and creative writing; projects emphasizing social-emotional development and encouraging exploration and discovery in academic areas; and projects focusing on preparing parents to improve the education and development of their children (GAO, 1975, pp. 3-4).

In 1969 the United States Office of Education contracted with the Stanford Research Institute to collect appropriate data as part of a national Follow Through evaluation. Since July 1972, Abt Associates,Inc. has been analyzing those data and communicating the results in a series of reports. This paper is based upon work performed in the most recent of those reports (Stebbins, St.Pierre, Proper, Anderson, and Cerva, 1977) in which the primary question addressed was whether the various educational strategies (operationalized through sponsors) being tested in Follow Through had differing impacts on the academic and affective levels of the pupils they served.*

* The data and results reported in this paper are a subset of the data and results included in the report by Stebbins, St.Pierre, Proper, Anderson and Cerva (1977). The interpretations placed on these data are intended to illustrate the way in which corrections for the unreliability of covariates change the conclusions of an evaluator, and are not meant to reflect the interpretations placed on the data by the Abt Associates evaluation team.

Method

The data on which this paper is based include measures on over 5000 children who began their Follow Through experience at entrance to kindergarten in the fall of 1971 and left Follow Through at exit from third grade in the spring of 1975. These pupils were distributed across nine sponsors, where each sponsor implemented its educational program in between five and seven school districts and where each school district contained a Follow Through treatment group (FT) and a non-Follow Through comparison group (NFT).

Sponsor effectiveness was judged in terms of both academic and affective outcomes and all children in the evaluation sample were administered the Metropolitan Achievement Tests (Elementary Level), the Raven's Progressive Matrices (modified version), the Coopersmith Self-Esteem Inventory, and the Intellectual Achievement Responsibility Scale at the end of third grade. These four tests contain 11 outcome scores which were grouped into three outcome domains as indicated in Figure 1.

Figure 1

DOMAINS OF THIRD GRADE TESTING IN FOLLOW THROUGH

Outcome Domain	Test	
BASIC SKILLS	Word Knowledge	Metropolitan Achievement Tests
	Spelling	
	Language	
	Math Computations	
COGNITIVE/ CONCEPTUAL SKILLS	Reading	
	Math Concepts	
	Math Problem Solving	
	Raven's Progressive Matrices	
AFFECTIVE OUTCOMES	Coopersmith Self-Esteem	
	Achievement Responsibility, Positive	Intellectual Achievement Responsibility Scale
	Achievement Responsibility, Negative	

The Basic Skills are the simplest objectives of traditional elementary schooling: vocabulary, spelling, the conventions of written language, and simple arithmetic computation. Cognitive/Conceptual Skills -- comprehension, reading, mathematical concepts, mathematical problems, and abstract problem-solving -- are also traditional academic goals, but are more complex and tend to require application of some basic skills. Affective Outcomes are approximate measures of the children's self-concept and of their tendency to attribute success and failure to themselves rather than to others. In addition, all pupils were administered a pretest, the Wide Range Achievement Test, upon entry to the program, and a set of standard student background measures were collected via parent interviews and school records.

The primary technique for isolating and strengthening the signal of the Follow Through effect from the noise in the data was a statistical adjustment of outcome scores based on preexisting conditions. The set of covariates included the pretest, first language (English vs. non-English), family income, highest occupation in household, ethnic membership (two vectors, White vs. other, Black vs. other), sex, entry age, and missing data codes (dummy variables coded 1 if missing and 0 if present) for income and occupation. In addition to these 10 variables, site specific (between site) covariates were coded for each sponsor to adjust for differences among sites. These between site covariates attempted to control for all nontreatment differences among children related to differences in the sites where the Follow Through experiment was implemented. An analysis of covariance was performed within each Follow Through sponsor for each of the 11 outcome measures. Differences among children related to the 10 covariates were adjusted out of each outcome measure with differences related to variations among sites within a particular sponsor being simultaneously controlled. The treatment condition was considered to be nested within each site and an adjusted outcome difference was estimated for each outcome within each site.

As stated earlier, the application of covariance techniques assumes that all covariates are perfectly reliable. However, such reliability cannot necessarily be assumed for each covariate in the present study.

Variables such as sex, ethnicity, income, occupation, education, language and age were all presumably measured with minimal error. The pretest posed the most serious problem. The reliability of the pretest was estimated on various Follow Through samples by a measure of internal consistency (coefficient alpha) and was on the order of .90.

Although there are several methods for dealing with a single fallible covariate (Porter and Chibucos, 1974), the solution to the problem in the multiple covariate case (even if only one of the covariates is unreliable) is not clear. Cohen and Cohen (1975) offer a method that has not been mathematically proven and which "rests on no more than the judgment of the present authors and some of our colleagues" (Cohen and Cohen, 1975, p. 373). Applying their method to the present case entailed correction only for the effects of unreliability in the pretest. The procedure involved correcting the correlations of the unreliable covariate with each other covariate and the outcome for attenuation due to unreliability by dividing each correlation by the square root of the estimated reliability of the covariate. In addition, the covariate standard deviation was corrected by multiplying the observed standard deviation by the square root of the covariate reliability.

There is disagreement in the literature as to the most appropriate measure of reliability to employ in such correction methods. Although the internal consistency (a statistic recommended as the appropriate measure of reliability by some methodologists) of the pretest was high (.9), Campbell and Boruch (1975) suggest that, as the time lapse between pretest and posttest increases, the correlation between them decreases. Consequently, they recommend the pre-post correlation be used as the appropriate measure of reliability.

Given this disagreement and the fact that a direct measure of the pre-post correlation for the pretest was not available, the Follow Through data were analyzed using three separate values spanning the range of potential estimates. The reliability values selected were .6, .8, and 1.0, the latter value being the equivalent of not correcting for unreliability.

With the child as the unit of analysis, the analysis estimated a set of raw score regression weights b_i for each sponsor and outcome using the model

$$\hat{Y} = a_o + \sum_{i=1}^{2s+9} b_i x_i$$

where a_o is a constant, s is the number of sites in a given sponsor, and $b_1 \ldots b_{2s+9}$ are the regression weights for the predictor variables $x_1 \ldots x_{2s+9}$ which are defined as follows:

$x_1 \ldots x_{10}$	= 10 covariates defined earlier
$x_{11} \ldots x_{s+9}$	= s-1 between site codes reflecting membership in the sponsor's sites (see Cohen and Cohen, 1975, pp. 171-211, for details on the coding of categorical variables with s distinct levels)
$x_{s+10} \ldots x_{2s+9}$	= s treatment within site codes

With the regression coded in this fashion, the s regression weights $b_{s+10} \ldots b_{2s+9}$ are interpretable as adjusted estimates of the FT/NFT outcome differences in the s sites. Thus, a total of 539 within-site estimates of FT effectiveness were calculated -- 11 estimates (one for each outcome) for each of 49 sites (nested within nine sponsors) in the analysis.

Results

Due to the complexity of the evaluation and the fact that a large number of adjusted outcome differences were computed, a system was devised to handle the interpretation of these results. Each was placed in one of three groups:

- positive treatment effect -- the Follow Through group in this site performed better than expected on this outcome given the performance of a similarly disadvantaged comparison group. An adjusted outcome difference was considered to represent a positive treatment effect if it favored FT, was statistically significant ($p<.05$), and greater in absolute magnitude than .25 standard deviation of the raw outcome measure.

- null treatment effect -- there was no difference between the performance of the Follow Through and comparison groups on this outcome in this site. An adjusted outcome difference was considered to represent a null treatment effect if it was not a positive or negative treatment effect.

- negative treatment effect -- the Follow Through group in this site performed less well than expected on this outcome given the performance of a similarly disadvantaged comparison group. An adjusted outcome difference was considered to represent a negative treatment effect if it favored NFT, was statistically significant ($p<.05$), and greater in absolute magnitude than .25 standard deviation of the raw outcome measure.

Summaries of the results of the three analyses categorized in the above fashion are presented at an aggregate level in Tables 1, 2, and 3 and indicate that across all sites, sponsors, and outcomes, lower pretest reliability estimates lead to movement of treatment effects from the null category. Correction for unreliability in the pretest tends, in the aggregate, to make the treatment effects less favorable to Follow Through: without correction, 463 (86 percent) of the effects are either positive or null; this number drops to 411 (76 percent) when corrected for a .80 reliability estimate and to 388 (72 percent) when corrected for a .60 reliability estimate.

However, the point of the evaluation was to compare the effectiveness of sponsors, not to search for a Follow Through main effect. In order to facilitate sponsor comparisons the treatment effects (classified as positive, null or negative) were aggregated by sponsor (nine sponsors) and outcome domain (shown earlier in Figure 1). The nine sponsors were each placed in one of three broad groups according to their areas of primary interest (see Figure 2). Such a categorization is not intended to reflect all the complexities and nuances of each sponsor's program. Readers interested in a description of each sponsor's program are referred to a report by Stebbins, Bock and Proper (1977).

Treatment effects were then aggregated by outcome domain within sponsor, and average sponsor treatment effects were calculated by assigning values of "1" to a positive treatment effect, "0" to a null treatment effect, and "-1" to a negative treatment effect. Figures 3, 4 and 5 present sponsor average treatment effects in each of the three outcome domains for the three different analyses while Table 4 presents the same data in tabular form.

Table 1

SUMMARY OF CHANGES IN TREATMENT EFFECTS BETWEEN UNCORRECTED ANCOVA AND ANCOVA WHEN PRETEST IS CORRECTED USING A RELIABILITY ESTIMATE OF .8

Corrected ANCOVA
(rel=.8)

		positive	null	negative		
Uncorrected ANCOVA (rel=1.0)	positive	32	0	0	32	
	null	19	360	52	431	percent agreement = 87
	negative	0	0	76	76	correlation = .76
		51	360	128	539	

Table 2

SUMMARY OF CHANGES IN TREATMENT EFFECTS BETWEEN UNCORRECTED ANCOVA AND ANCOVA WHEN PRETEST IS CORRECTED USING A RELIABILITY ESTIMATE OF .6

Corrected ANCOVA
(rel=.6)

		positive	null	negative		
Uncorrected ANCOVA (rel=1.0)	positive	29	3	0	32	
	null	27	324	80	431	percent agreement = 79
	negative	0	5	71	76	correlation = .63
		56	332	151	539	

Table 3

SUMMARY OF CHANGES IN TREATMENT EFFECTS BETWEEN ANCOVA WHEN PRETEST IS CORRECTED USING A RELIABILITY ESTIMATE OF .8 AND ANCOVA WHEN PRETEST IS CORRECTED USING A RELIABILITY ESTIMATE OF .6

Corrected ANCOVA
(rel=.6)

		positive	null	negative		
Corrected ANCOVA (rel=.8)	positive	44	7	0	51	
	null	12	314	34	360	percent agreement = 88
	negative	0	11	117	128	correlation = .81
		56	332	151	539	

Figure 2

FOLLOW THROUGH MODELS BY PRIMARY EMPHASIS

PRIMARY EMPHASIS	SPONSOR/MODEL NAME
Basic Skills - These models focus first on the elementary skills of vocabulary, arithmetic computation, spelling, and language.	● University of Oregon - Direct Instruction Model ● University of Kansas - Behavior Analysis Approach ● Southwest Educational Development Laboratory - Language Development Education Approach
Cognitive/Conceptual - These models emphasize the more complex "learning-to-learn" problem solving skills.	● University of Florida - Florida Parent Education Model ● Arizona Center for Early Childhood Education - Tucson Early Education Model ● High/Scope Educational Research Foundation - Cognitively Oriented Curriculum Model
Affective/Cognitive - These models focus primarily on self-concept and attitudes toward learning, and secondarily on "learning-to-learn" skills.	● Far West Laboratory for Educational Research and Development - Responsive Education Model ● Bank Street College of Education - Bank Street College of Education Approach ● Education Development Center - EDC Open Education Follow Through Program

FIGURE 3:

Sponsor Average Treatment Effects in Basic Skills

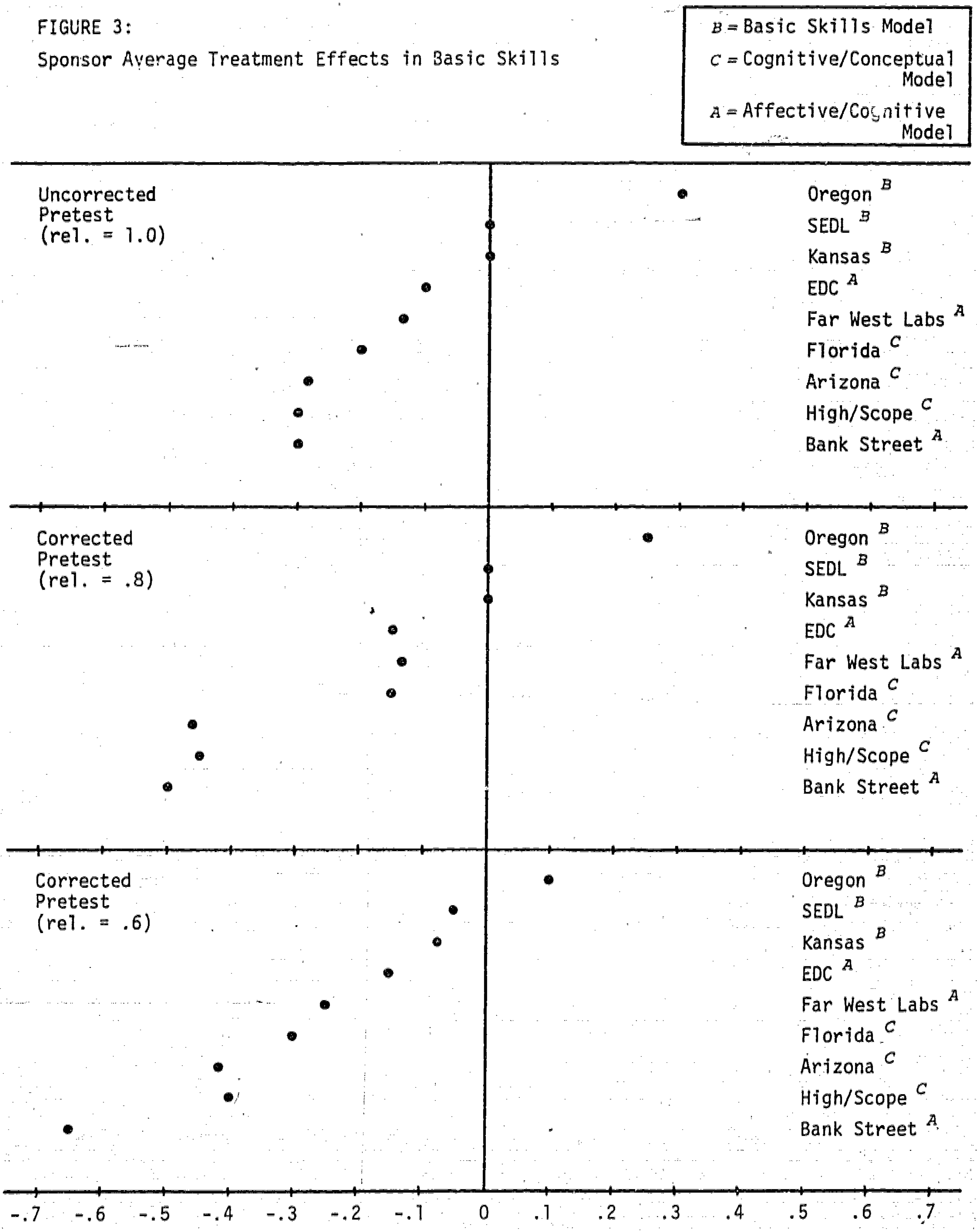


FIGURE 4:

Sponsor Average Treatment Effects for Cognitive/Conceptual Skills

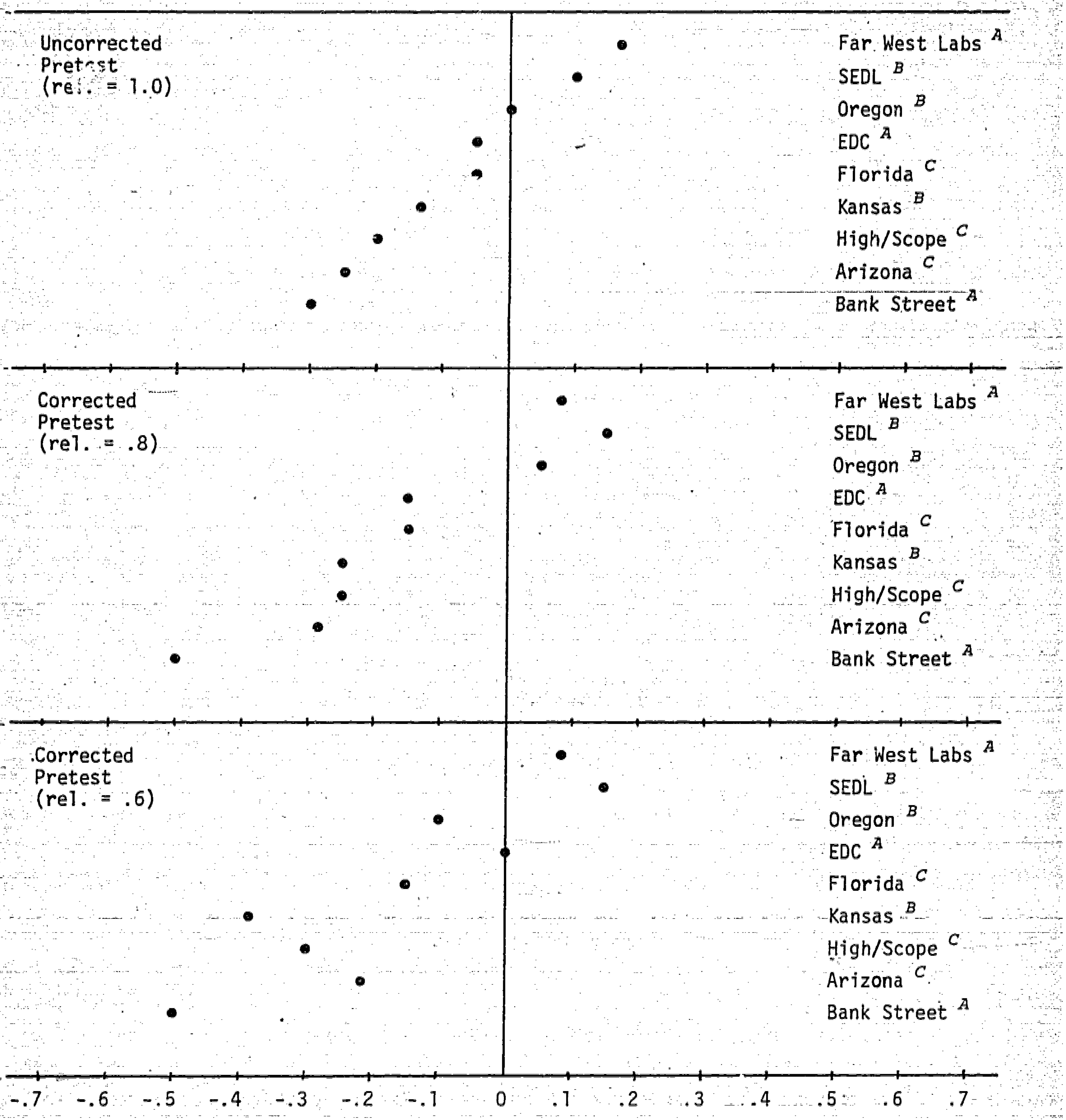


FIGURE 5:

Sponsor Average Treatment Effects in Affective Outcomes

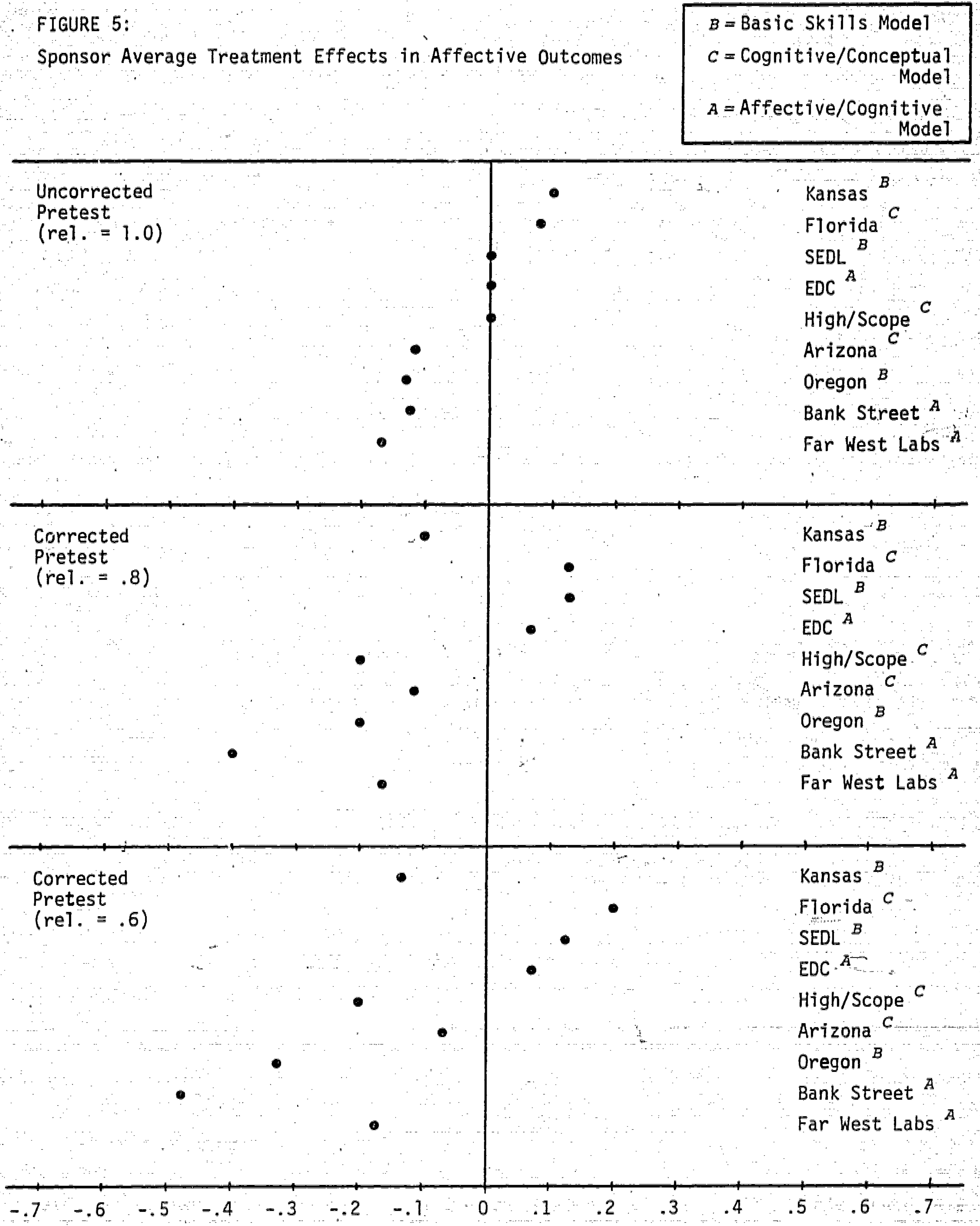


Table 4

SPONSOR AVERAGE TREATMENT EFFECTS IN BASIC SKILLS, COGNITIVE/CONCEPTUAL SKILLS AND AFFECTIVE OUTCOME AREAS

Sponsor	BASIC SKILLS			COGNITIVE/CONCEPTUAL SKILLS			AFFECTIVE OUTCOMES		
	uncorrected rel = 1.0	corrected rel = .8	corrected rel = .6	uncorrected rel = 1.0	corrected rel = .8	corrected rel = .6	uncorrected rel = 1.0	corrected rel = .8	corrected rel = .6
Oregon[B]	.30	.25	.10	.00	.05	-.10	-.13	-.20	-.33
Kansas[B]	.00	.00	-.07	-.14	-.25	-.39	.10	-.10	-.14
SEDL[B]	.00	.00	-.05	.10	.15	.15	.00	.13	.13
Arizona[C]	-.29	-.46	-.42	-.25	-.29	-.21	-.11	-.11	-.06
High/Scope[C]	-.30	-.45	-.40	-.20	-.25	-.30	.00	-.20	-.20
Florida[C]	-.20	-.15	-.30	-.05	-.15	-.15	.07	.13	.20
Far West Labs[A]	-.13	-.13	-.25	.17	.08	.08	-.17	-.17	-.17
Bank Street[A]	-.30	-.50	-.65	-.30	-.50	-.50	-.13	-.40	-.47
EDC[A]	-.10	-.15	-.15	-.05	-.15	.00	.00	.07	.07

[B] Basic Skills Model

[C] Cognitive/Conceptual Model

[A] Affective/Cognitive Model

An examination of these data reveals some interesting overall patterns. First, correction for unreliability in the pretest appears to distort the rank order of the sponsors less with respect to Basic Skills than with respect to Cognitive/Conceptual or Affective outcomes. Second, such corrections tend to produce lower estimates of the absolute level of sponsor effectiveness in all three outcome areas. This is most pronounced in Basic Skills. At a less global level it can be seen that according to the analysis using an uncorrected pretest (rel = 1.0), the models which emphasize Basic Skills do better on tests of these skills than models which emphasize the Cognitive/Conceptual or Affective areas. In particular, the University of Oregon's Direct Instruction Model is clearly more effective in Basic Skills than the rest. Correction of the pretest does little to alter this interpretation when a reliability coefficient of .8 is assumed: Oregon still appears to perform best and the Basic Skills models have higher average treatment effects than other model types. Note though, that sponsors in general have lower estimated levels of effectiveness in Basic Skills when the pretest is corrected (rel = .8). Changing to a reliability estimate of .6 further depresses overall averages, but does little to alter the relative standing of sponsors.

An examination of average sponsor treatment effects in Cognitive/Conceptual Skills reveals a somewhat different pattern. When the pretest is not corrected Far West Labs is the best performer and no single model type appears most effective. Correction of the pretest for unreliability (rel. = .8) changes this interpretation slightly as the estimate of SEDL's effectiveness is raised while Far West Labs becomes less effective. Use of a reliability estimate of .6 further alters the relative standing of some sponsors, although none change more than one or two rank positions. As is the case with Basic Skills outcomes, most sponsors appear less effective in terms of Cognitive/Conceptual Skills when the pretest is corrected.

With respect to the Affective area, Kansas and Florida are the most effective sponsors in the uncorrected analysis (rel = 1.0). Correction of the pretest (rel = .8) dramatically lowers the estimate of effectiveness for Kansas while raising it for Florida, SEDL and EDC.

Changing to an estimate of .6 further separates the sponsors. Again, the overal_ effect of correcting the pretest for assumed unreliability is to lower our estimate of effectiveness for most sponsors.

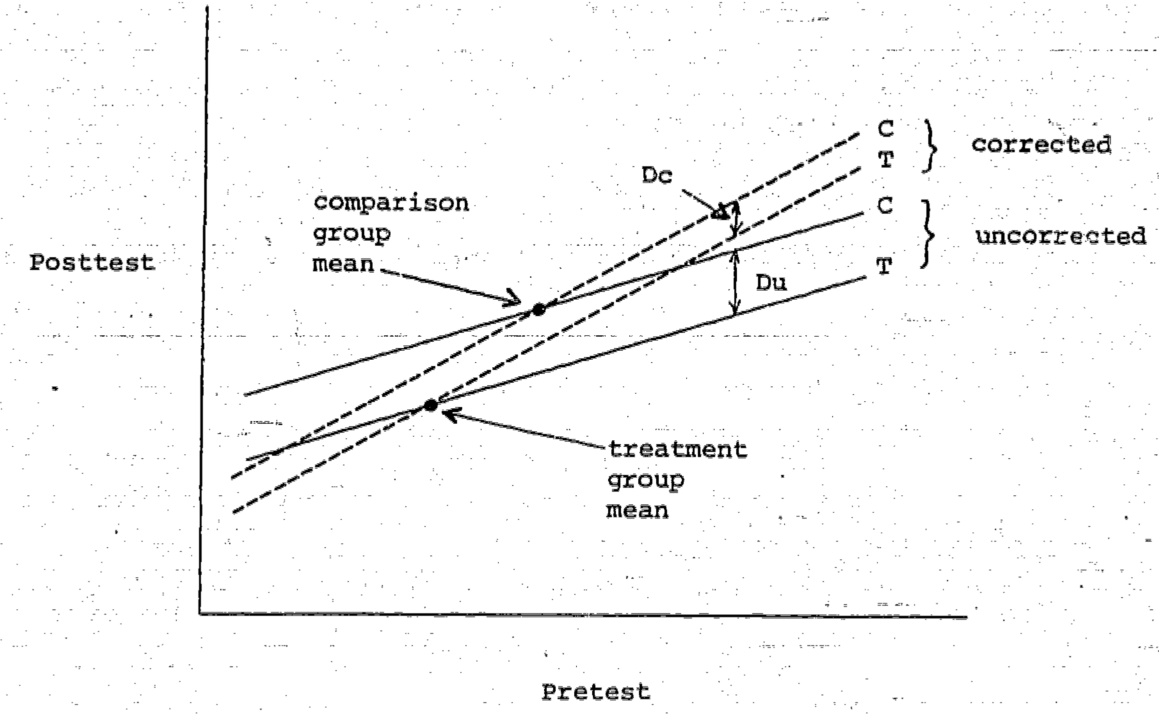
The application of pretest corrections, therefore, changes the estimates of both the relative standing of sponsors and the absolute level of sponsor effectiveness differentially by outcome area. The changes in interpretation are clearest in Basic Skills, where the ranking of sponsors is essentially preserved, and the overall level of effectiveness is lowered fairly consistently across sponsors. The same pattern is less evident but still present in the Cognitive/Conceptual and Affective areas where changes in the rank order of sponsors occur more often.

Discussion

As noted in the results section, the primary effect of correcting the pretest for assumed unreliability is to deflate the uncorrected estimates of Follow Through effectiveness while essentially preserving the rank order of sponsors. This pattern is more clearly seen with respect to Basic Skills than other outcomes. The question which now arises is: Why did this happen? Let us first consider what might occur when a covariate is corrected for unreliability in the evaluation of a typical compensatory education program. In such a program we expect the treatment group to have a lower pretest mean than the comparison group, and therefore, correction for unreliability in the pretest should act to make adjusted posttest differences more favorable to the treatment group. Figure 6 shows an example where the treatment group mean score is below that of the comparison group on both the pretest and posttest. The parallel solid lines represent the regression lines for the treatment and comparison groups in an uncorrected analysis while the dashed lines represent the regression lines for the same groups when the pretest has been corrected for unreliability. Since the regression lines must pass through the means of their samples, and the slope of the regression lines in the corrected analysis is, by definition, steeper than that in the uncorrected analysis, the separation of the regression lines and hence the adjusted mean difference between the treatment and comparison groups is smaller for the analysis using the corrected pretest ($D_c < D_u$) -- correction for unreliability has improved the standing of the treatment

Figure 6

REGRESSION LINES FOR TREATMENT AND COMPARISON GROUPS IN A CORRECTED AND UNCORRECTED ANALYSIS



group. Now, the effect of this correction depends on the location of the treatment and comparison group means and on the pre-post correlation, and Figure 6 is only meant to represent a single situation -- one which is likely to occur in the evaluation of compensatory education programs.

Let us see if an examination of pretest means for the Follow Through sponsors allows us to apply the above logic to Follow Through. Table 5 presents descriptive statistics by treatment group within sponsor for the pretest and the four Basic Skills posttests. It can be seen that the treatment group (FT) scores substantially lower than the comparison group (NFT) in only two sponsors, Arizona and Far West Labs. For all other sponsors the FT group scores above or about the same as the NFT group. This suggests that Arizona and Far West Labs might appear more effective when the analysis is corrected for unreliability in the pretest while other sponsors would appear less effective or show no change. However, this is not the case. A reexamination of Figures 3, 4 and 5 shows that these two sponsors do not gain in effectiveness in the corrected analyses. In fact, the overall pattern of sponsors appearing less effective in the corrected analyses is very strong, a finding which is not intuitively appealing. It would seem that sponsors with treatment groups that score lower than comparison groups on the pretest should be helped by correction for unreliability. Perhaps there is some other factor operating which is causing the general drop in program effectiveness.

Table 6 presents adjusted outcome differences (regression weights for the FT/NFT within-site contrast -- corresponding to variables x_{s+10} ... x_{2s+9} in the analytic model presented earlier), associated standard errors, and t-ratios by outcome for the uncorrected and the two corrected analyses. The data in this table are averages of statistics calculated for each site within each sponsor. There are 49 sites in the nine sponsors, and therefore each number presented in Table 6 is based on 49 site level pieces of data. It can be seen that across analyses there is very little change in the average adjusted outcome differences. On the other hand, there is a pronounced reduction in the size of the standard errors of those adjusted differences (on the order of a 30 percent decrease between the standard error of the uncorrected and corrected for rel = .8 analyses). These two conditions lead to an increase in the

Table 5

DESCRIPTIVE STATISTICS FOR THE PRETEST AND BASIC SKILLS OUTCOMES BY SPONSOR AND TREATMENT GROUP

					Basic Skills Outcomes							
			Pretest		Word Knowledge		Spelling		Language		Math Computations	
Sponsor	Treatment	N	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Oregon	FT	316	29.6	10.3	24.6	9.2	20.2	11.6	21.2	10.0	22.5	8.7
	NFT	317	30.9	12.9	27.8	11.2	23.0	12.5	19.2	9.6	19.9	7.8
Kansas	FT	585	28.2	11.6	24.0	10.5	19.9	12.6	16.9	8.5	19.2	7.6
	NFT	762	26.2	12.1	22.9	10.4	19.4	12.9	15.8	7.5	16.2	6.8
SEDL	FT	492	26.1	12.8	19.9	9.7	14.0	12.1	15.6	7.2	17.9	7.3
	NFT	563	26.6	11.1	21.4	9.5	17.9	12.6	15.1	6.9	15.6	6.5
Arizona	FT	329	30.6	13.1	24.1	10.8	17.4	11.3	17.6	8.5	18.6	7.8
	NFT	292	35.2	13.4	33.8	11.1	25.4	11.2	22.7	10.0	22.1	8.3
High/Scope	FT	177	28.1	12.0	19.9	10.9	14.8	19.8	13.2	6.2	15.3	6.7
	NFT	337	29.4	12.3	24.9	10.8	12.8	13.0	16.9	8.0	17.6	7.4
Florida	FT	254	27.8	11.7	23.4	10.9	17.1	12.2	15.8	7.1	15.9	15.7
	NFT	481	27.2	12.1	22.2	10.9	18.4	12.8	16.0	8.3	6.0	6.8
Far West Labs	FT	241	28.9	12.4	22.7	11.5	15.1	12.5	15.7	7.6	17.3	7.1
	NFT	277	32.2	12.2	27.0	11.3	20.3	12.3	18.7	8.7	18.2	7.1
Bank Street	FT	264	31.3	12.8	23.5	10.9	17.5	12.5	16.9	8.5	16.4	6.8
	NFT	587	28.3	12.0	24.0	10.7	20.8	12.7	16.2	8.3	16.9	7.2
EDC	FT	248	28.7	12.0	21.8	11.4	15.6	12.4	16.4	8.6	17.1	6.4
	NFT	487	29.1	12.6	23.5	11.0	19.6	12.8	16.7	9.0	16.8	7.5

Table 6

AVERAGE ADJUSTED OUTCOME DIFFERENCE, STANDARD ERROR AND T-RATIO BY OUTCOME FOR UNCORRECTED AND CORRECTED ANALYSES

			ANALYSIS								
			Uncorrected (rel = 1.0)			Corrected (rel = .8)			Corrected (rel = .6)		
Outcome Domain	Outcome	N	Adj. Diff.	Std. Error	t-Ratio	Adj. Diff.	Std. Error	t-Ratio	Adj. Diff.	Std. Error	t-Ratio
Basic Skills	Word knowledge	49	-2.03	2.50	-.93	-2.06	1.74	-1.34	-2.07	1.50	-1.62
	Spelling	49	-2.56	3.08	-.98	-2.61	2.18	-1.41	-2.91	1.96	-1.75
	Language	49	-.69	1.99	-.39	-.73	1.41	-.54	-.78	1.26	-.69
	Math Computations	49	.38	1.80	.28	.36	1.29	.37	.32	1.19	.35
Cognitive/ Conceptual Skills	Ravens	49	-.47	1.17	-.46	-.48	.83	-.73	-.51	.77	-.86
	Reading	49	-.89	1.80	-.55	-.94	1.25	-.82	-.93	1.13	-.96
	Math Concepts	49	-.71	1.78	-.42	-.73	1.24	-.63	-.76	1.07	-.80
	Math Probl. Solv.	49	-.19	1.59	-.18	-.20	1.13	-.30	-.23	1.02	-.38
Affective Outcomes	Coopersmith	49	-.92	2.08	-.51	-.93	1.53	-.72	-.95	1.50	-.76
	IARS (-)	49	-.06	.77	.01	-.07	.57	.02	-.07	.56	-.00
	IARS (+)	49	-.09	.68	-.17	-.09	.49	-.25	-.10	.48	-.30

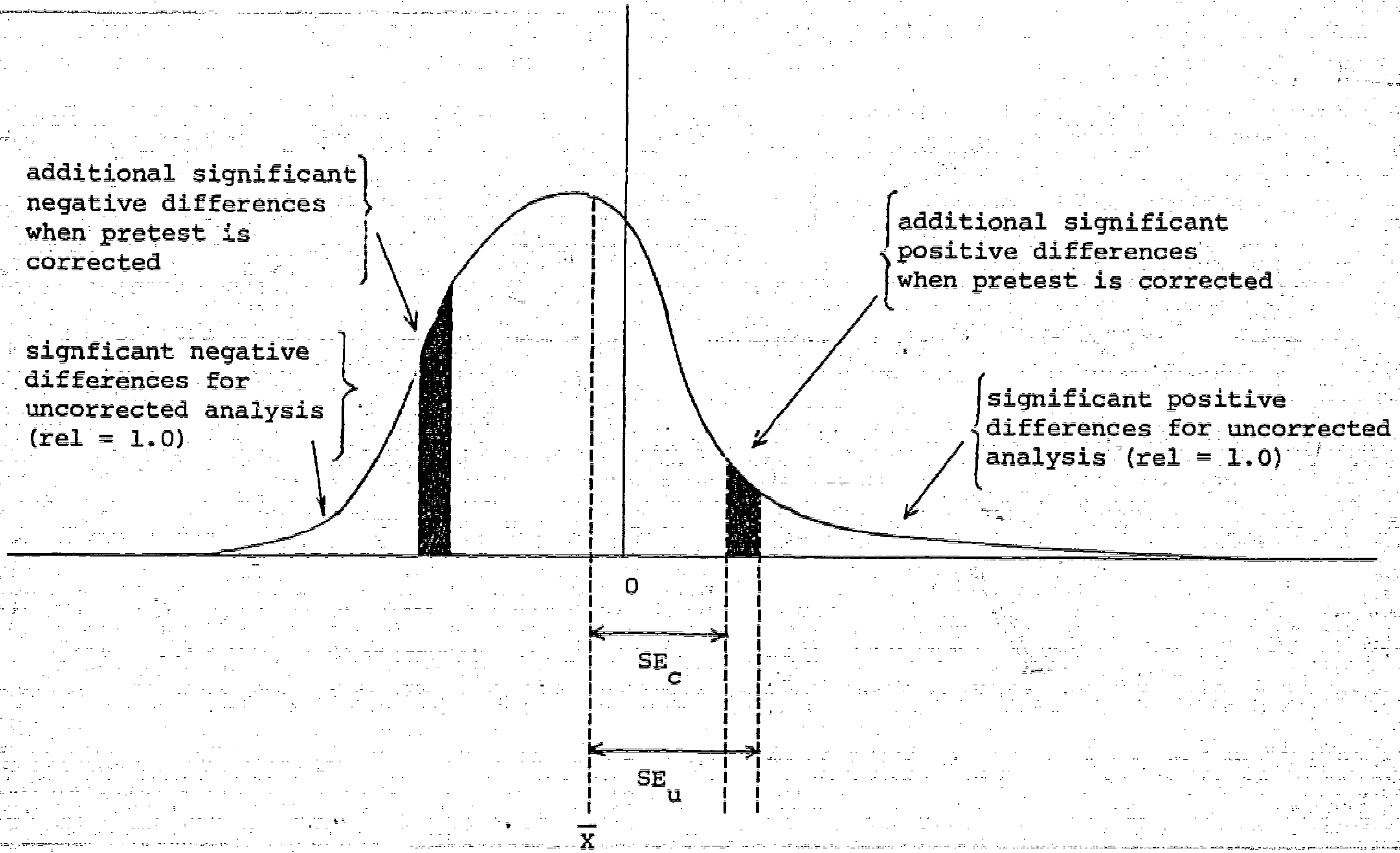
size of t-ratios, which are derived by dividing the adjusted outcome difference by its associated standard error,* and a corresponding increase in the number of significant effects. Since Table 6 shows that the distribution of adjusted outcome differences has a mean less than zero for all outcomes except math computations, and since these distributions tend to be positively skewed, the effect of reducing the standard error is to increase the number of negative effects at a faster rate than the number of positive effects. Figure 7 gives a representation of how this happens.

We have seen that correction of the pretest for assumed unreliability can lead to changes in the conclusions that an evaluator reaches in terms of the rank order of sponsors as well as the overall level of program effectiveness (across sponsors). Such changes were shown to be dependent on a variety of factors. First, Basic Skills outcomes, which are likely easiest to measure and hence the most reliable, show the fewest changes in rank order among sponsors, while Affective outcomes, surely the most difficult to measure and hence the least reliable, show the most changes in rank order among sponsors. Second, changes in conclusions do not depend directly on treatment/comparison group pretest differences. The two sponsors with treatment groups that scored lower than their comparison groups on the pretest did not particularly benefit from the application of corrections for unreliability. Third, changes in conclusions depend on the initial level of program success. To the extent that standard errors are lessened it becomes easier to find statistically significant differences between groups. Fourth, although not investigated in this paper, the existence of covariates other than the pretest can have an important effect on the results since the correlation of each other covariate with the pretest as well as the pretest/outcome correlation is corrected. Finally, the appropriateness of the pretest reliability coefficient must be considered. If the appropriate reliability is on the order of .90 as a coefficient of

* Note that the average t-ratio does not necessarily equal the average adjusted outcome difference divided by the average standard error.

Figure 7

REPRESENTATION OF THE DISTRIBUTION OF ADJUSTED OUTCOME DIFFERENCES FOR A GIVEN OUTCOME MEASURE


additional significant negative differences when pretest is corrected
signficant negative differences for uncorrected analysis (rel = 1.0)
additional significant positive differences when pretest is corrected
significant positive differences for uncorrected analysis (rel = 1.0)
0
SE_c
SE_u
$\overline{X}$

SE_c = two standard errors according to corrected analysis

SE_u = two standard errors according to uncorrected analysis internal consistency shows, correction for unreliability will make very little difference. On the other hand, the lower estimates of pretest reliability used in this study lead to increasingly important changes in conclusions.

BIBLIOGRAPHY

Campbell, D. T. and Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives. In C. A. Bennett and A. A. Lumsdaine (Eds.), Evaluation and Experiment, New York: Academic Press, 1975.

Cohen, J. and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

GAO. Follow Through: Lessons learned from its evaluation and the need to improve its administration (Report to Congress, MWD-75-34). Washington, D.C.: GAO, 1975.

Lord, F. M. Large-scale covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 307-321.

Porter, A. C., and Chibucos, T. R. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Stebbins, L. B., Bock, G., and Proper, E. C. Education as experimentation: A planned variation model, Volume IV-B. Cambridge, Massachusetts: Abt Associates, Inc., 1977.

Stebbins, L. B., St.Pierre, R. G., Proper, E. C., Anderson, R. B., and Cerva, T. R. Education as experimentation: A planned variation model, Volume IV-A. Cambridge, Massachusetts: Abt Associates, Inc., 1977.

Wolff, M. and Stein, A. Six months later, Head Start evaluation project. New York: Yeshiva University, Ferkauf Graduate School of Education, 1966.